ИНФОРМАТИКА

UDC 004.738.5:51-7
MSC 68R10

Application of webometrics methods for analysis and enhancement of academic site structure based on page value criterion

A. M. Nwohiri[1], *A. A. Pechnikov*[2]

[1] University of Lagos, University Road, Akoka, Yaba, Lagos, 101017, Nigeria
[2] Institute of Applied Mathematical Research of the Karelian Research Centre, Russian Academy of Sciences, 11, Pushkinskaya ul., Petrozavodsk, 185910, Russian Federation

For citation: Nwohiri A. M., Pechnikov A. A. Application of webometrics methods for analysis and enhancement of academic site structure based on page value criterion. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2019, vol. 15, iss. 3, pp. 337–352. https://doi.org/10.21638/11702/spbu10.2019.304

This paper describes a formalized procedure for exploring a site using webometrics methods. The procedure involves gathering details on a site's structure, constructing and exploring the resulting webgraph, defining the correctness criterion, identifying control actions that would improve the structure under the given criterion, testing the correctness criterion on real-world examples and developing recommendations on improving the structure. PageRank is used as a criterion to evaluate the value of web pages. The value is determined by the presence/absence of a link pointing to that page from the homepage of the site. Going by the correctness criterion, valuable pages of a site should have the highest PageRank among all other pages of that site. Control action consists of removing non-valuable directories (and transforming them into independent sites), whose root page has a high PageRank. Experiments are conducted on three faculty sites of major universities in USA, Russia and Nigeria. The approach is shown to be applicable and reasonable in all cases.

Keywords: website, graph, PageRank, universities, data mining, website structure, web harvesting, web mining, URL.

1. Introduction. This paper explores webometrics techniques with the aim of improving the structure of a website. These techniques are based on web graph analysis. Webometrics (also cybermetrics) is a branch of computer science dedicated to the study of the quantitative aspects of the construction and use of information resources, structures and technologies in relation to the web [1, 2].

The html pages of a site and the hyperlinks connecting those pages may be respectively viewed as nodes and arcs in a directed webgraph of that site. The webgraph is one of the most natural mathematical models of the web. Webgraph is a directed graph consisting of html pages (nodes) and hypertext links (arcs) connecting the pages. A webgraph could be examined "in general" at the initial stage of web development. A good example is the bow-tie structure by [3], where about two hundred million pages, indexed by the AltaVista search engine, were used as nodes. With rapid evolution in web science, websites (and their interconnected sets, which are fragments of the "big" web) have replaced web pages as structural units of research. However, such structures are well interpreted by graphs [4, 5].

Today, websites are used as complex and attractive tools for exploring web objects that are sometimes comparable to web fragments. For example, Google search engine was used to find about 40,000 web pages on the website of the University of Oxford (www.ox.ac.uk). It is therefore rational to apply known approaches and techniques, including the webgraph (as the model of site structure), in the study of a site structure.

Analysis of a webgraph as a website model has uncovered features, which, when interpreted on a real site, provide guidelines on how to structurally improve a site. This subject is well covered in [6, 7].

In this work, it is assumed that website design and maintenance is a goal-oriented process. It is also supposed that site owners are always thinking on how to improve their site without always clearly articulating the improvement criteria. Unlike many other web resources, the websites of business entities, research institutions, and universities can be referred to as the so-called "regulated web resources". Regulated means there are official regulations setting out the goals and objectives of such website, the main structural components, the rules for adding and modifying information on that site, etc. Such regulations should exist. However, most official websites were found not to have such regulations. To control some processes in the web, one would need to formulate managerial decisions via official documents. The methods by which such decisions are implemented ought to be outlined. Before the appropriate managerial decisions are implemented, information on the particular subject area (in this particular case — website) is gathered. This eventually leads to construction of the formal models of the subject area, formulation and solving of optimization problems. One must define the potentially possible correctness criteria for the site and the so-called "control actions". Recommendations on how to boost the site under such criteria must be developed. Control actions are treated in the subsection "Determining the Correctness Criterion for a Website Structure".

Below is a brief description of the mathematical modeling process with regards to a website as the subject area.

Step 1: Gathering some data about the website structure.

Step 2: Using the gathered data to construct and explore a webgraph for the site.

Step 3: Defining a possible correctness criterion for the website structure.

Step 4: Outlining control actions that would improve the site structure under the given correctness criterion.

Step 5: Coming up with recommendations and guidelines on how to enhance the site structure.

The above steps are applied on specific websites of faculties from three selected major universities in Nigeria, Russian Federation and USA (see section 4).

To the authors, it is submitted that the discovered properties of these sites — relatively huge number of web pages, content diversity, sufficiently independent fragments, and others — provide a platform for explaining (through simple and understandable examples) all the steps involved in the research.

The proposed procedure comes with high level of generality. It can be applied to the study of many websites. Obvious constraints are down to the actual sizes of sites, features of the software used and the power of computing resources.

2. Background. The study of the web as a graph and the analysis of hyperlink structure are fairly new areas of research. Research on webgraph structure gives valuable insight into the social mechanisms governing the evolution of structure. Such studies are important for creating better crawling algorithms, designing ranking techniques, and building clear-cut web structure models.

Experiments on the local and global properties of a webgraph were presented in [3]. These properties were revealed using the data harvested by Altavista search engine. The authors claim that the macroscopic structure of the web is significantly more complicated than suggested by earlier experiments on a smaller scale.

Link relationships in the Nordic academic web space, comprising of 23 Finnish, 11 Danish and 28 Swedish academic web domains were investigated by [4]. It was discovered that the Nordic network is made of a cohesive network of three clearly defined sub-networks. It was also uncovered that this Nordic network rests on the Finnish and Swedish sub-networks.

Page categorization was applied in [5] to show that one can produce even stronger associations by restricting the metrics to subsets that are more closely connected to the research of the host university. The authors found that there was a partial overlap between the effects of applying advanced document models and separating page types. However, it was argued that combination of the two achieved the best results.

A new technique for evaluating and improving website link structure (by using web usage mining) and assessing the usage pattern of an Iranian organization, was outlined in [6]. The study found that the use of graph theory in assessing website usability has some advantages over other methods of website usability and structure analysis.

A new class of processes "the web Markov skeleton processes" (WMSP) arising from information retrieval on the web, were proposed and discussed by [7]. The paper examined the scope and time homogeneity of WMSPs, and explored a new class of processes, the so-called mirror semi-Markov processes. Some WMSPs applications in computing the importance of a page on the web were briefly reviewed.

In [8], webometrics techniques were, for the first time, deployed to analyze the websites of the academic institutes of the Siberian Branch of the Russian Academy of Sciences (SB RAS). The researchers used Google, Yahoo and Yandex search engines to uncover the quantitative characteristics of the sites. Out of the 80 sites (institutes) studied, about 39 % of them were found to have small number of pages (less than 100 pages), while 24 % have very large number of pages (above 1000 pages). It was also established that 12 institutes have over 1000 external links, while 20 institutes (25 %) have just less than 100 external links.

Two webspaces — one belonging to the academic institutions of SB RAS, and the other belonging to Fraunhofer—Gesellschaft, a German research organization — were investigated in [9]. The scientific communities of the underlying websites were examined. A mathematical model for the academic webspaces of the two organizations was presented. A hypothesis that communities in these academic networks should reflect scientific

collaborations between the corresponding institutes was proposed and checked for SB RAS.

A large webgraph comprising of over 3.5 billion web pages and 128.7 billion hyperlinks was analyzed in [10]. Features such as degree distributions, connectivity, average distances, and the structure of weakly/strongly-connected components were analyzed and compared. Results showed that some of the features previously observed in [3] were very dependent on artefacts of the crawling process, whereas others appear to be more structural. However, the authors found that very different proportions of nodes can reach or can be reached from the discovered giant strongly connected component, suggesting that the "bow-tie structure" — as described in [3] — is strongly dependent on the crawling process. It was thus concluded that the "bow-tie structure" is not a structural property of the web.

3. Methods. A website consists of html pages and web documents linked to each other by internal hyperlinks. This collection is unique in content and identified on the web with a unique domain name.

Based on the above definition, the "site/sub-site" and "domain/sub-domain" relationships are not used in this research. For example, the Faculty of Applied Mathematics and Control Processes (apmath.spbu.ru) is under Saint Petersburg State University (www.spbu.ru). However, both have different sites. So in this context, both sites are considered as different sites in this paper.

This approach is justified. There are many cases where the parent website of an organization is developed and administered by one group of employees, while the sites of departments, conferences, and journals falling under the same organization, are administered by another group. In such cases, it is safe to say that the sites do not have unified goals or common creation and functioning approaches and methods.

3.1. Generating the structural summary of a site. To build a website model, details about html pages and site documents, and of course the internal hyperlinks connecting them are needed.

Web crawlers are used to retrieve data from the web. These crawlers scan through web pages and/or web documents to harvest certain information, identify patterns, create statistics or preserve site resources [11]. For example, web spider engine JSPider (http://jspiders.com) reviewed by [12], basically deals with construction of site structure. There are so many other publicly accessible web crawlers that can be deployed to scan a website.

However, for some important reasons, the authors in this paper preferred using their own web crawler called Rapid Configurable Crawler (RCCrawler) [13]. One of such important reasons is the problem of dust — a situation whereby different URLs have similar text. As pointed out by [14, p. 1], "Such duplicate URLs are prevalent in web sites, as web server software often uses aliases and redirections, translates URLs to some canonical form, and dynamically generates the same page from various different URL requests".

The ability to detect and remove dust from previous crawl logs enhances crawling efficiency, reduces crawling time and improves the credibility of popularity indicators. In other publicly accessible web crawlers, the authors do not know how the dust problem is resolved and whether it was even resolved. In cases where the dust problem is not resolved, using such a crawler would introduce "extra" nodes in the webgraph, thereby corrupting and complicating the graph model. The dust problem is resolved in RCCrawler.

Investigations have shown that new tasks and functions emerge when gathering and interpreting information. Thus, constant improvement in crawling capabilities has been made necessary. Lack of documentation complicates the process of implementing such tasks

and functions. So another reason why RCCrawler was chosen consists of the differences in research approaches and problem statements. This is reflected in the algorithms and structures of hyperlink databases. In addition, the choice of a particular crawler largely depends on crawling convenience and on credibility of obtained results.

RCCrawler is a cross-platform application framework written in C++ (C++11 standard) using a set of libraries and an additional preprocessor. It comes with a common architecture that can be instantiated by certain classes and, thus, implement various crawling tasks. Detailed description of the structure, technical features, specific tasks and usage examples of RCCrawler are presented in [13].

The web robot systematically browses a site, starting from the home (index) page and navigating via internal hyperlinks in a given traversal order. This is known as the breadth-first search [15]. The required scan depth can be set. After scanning a site, the crawler generates SCV files containing lists of internal hyperlinks connecting the files.

3.2. Webgraph construction and investigation. Information about html pages, web documents and internal hyperlinks harvested by RCCrawler is used to construct the webgraph of that particular site. The nodes of this webgraph are the html pages and web documents found, while the arcs are the hyperlinks discovered.

A webgraph has three important and pretty obvious properties:

- the webgraph of a site has a dedicated source node — the homepage (index page) of that site;
- the level structure of the webgraph of a site can be determined using the distance of any web page from the homepage;
- the webgraph of a site is always a connected graph and is almost always a strongly connected graph.

In graph theory, the level of the source node of a webgraph is 0. The level of nodes that can be accessed from the source node in one step is 1. The level of nodes that can be accessed from the source node in two steps is 2, etc.

The connectivity property is inherent in any webgraph built both on a set of html pages and on a set of html pages and web documents. This ensues from the way a site is traversed — lists of all subsequent pages are generated when analyzing the previous ones, starting with the homepage. Obviously, nodes representing web documents are the leaf nodes — have incoming but no outgoing arcs. Not all webgraphs are strongly connected. For example, a graph containing leaf nodes is not a strongly connected graph. A webgraph containing only a set of html pages is almost always a strongly connected graph because any page in the site contains at least one link pointing to the home page. Such a single link could be the logo of the organization pasted on each page. Figure 1 shows an example of a webgraph. The root node (homepage) is designated as page0. Nodes page1, page2, and page3 are level 1 nodes (pages with direct links to/from the homepage).

Gephi — a visualization software system for all kinds of graphs and networks [16] — was used to analyze the webgraphs of sites. This program visualizes and draws graph. It also calculates basic characteristics, such as node degree, graph diameter, graph density, modularity, PageRank, connected components and other indicators.

3.3. Determining the Correctness Criterion for a Website Structure. This section deals with one of the most important methodological problems in webometrics: what is the real purpose of creating a website? As mentioned earlier, when the rules governing a site are known, one can formulate criteria that tell whether implementation of a site complies with requirements stated in relevant regulations.

In most cases known to the authors, there are no such regulations. So, the authors try

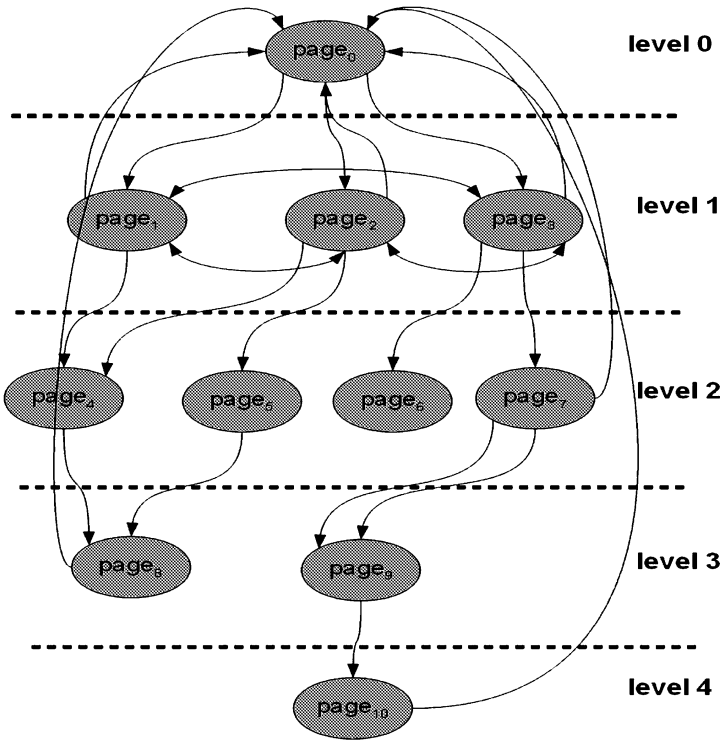

Figure 1. Webgraph of a website

to formulate a plausible correctness criterion for a site structure. The criterion will depend on common sense, the author's web development experience, and experiments conducted on a large set of real websites.

One of the most famous tools used in estimating the value or quality of pages on the web is the PageRank [17]. The authors use the PageRank (PR) to determine how important the pages of a website are. This is based only on the site's internal structure. It involves calculating the PR of the webgraph nodes (pages) of the site.

In a perfect case, the following could be the correctness criterion for a website structure: for web developers, pages with the highest PRs are the most "valuable" pages of a site.

However, the situation considered in this paper is a non-perfect case. Therefore, the most valuable web pages need to be defined. It is assumed that the homepage of a site and other pages corresponding to major sections on that site are considered as valuable pages by web developers. In this research, nodes of a graph representing valuable pages are called valuable nodes.

The valuable web pages of a particular site are identified as follows:

Step 1: The site in question is scanned using RCCrawler.

Step 2: With the data harvested by RCCrawler, the webgraph of that site is constructed.

Step 3: The PR value of each node (page) of the constructed webgraph is calculated.

Step 4: The calculated PR values are arranged in descending order.

Step 5: The arranged PR vector is plotted against the number of pages (Figure 2).

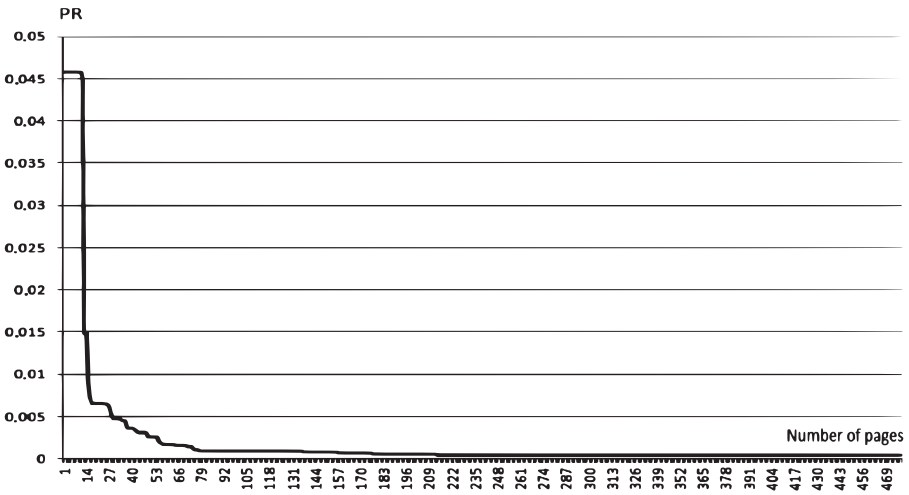

Figure 2. Graph of PR values of webgraph nodes for cmns.umd.edu

Such a graph (PR distribution) is characteristic of any webgraph of a modern website — valuable nodes (pages) typically have the highest PR values and such nodes are not much in number [18].

From the above-stated, it can be concluded that valuable nodes with low PRs and non-valuable nodes with high PRs violate the correctness criterion.

So for a given website, the procedure for detecting any violations of the correctness criterion can be described as follows:

Step 1: On the homepage of that website, all the headings of major sections (pointing to corresponding pages) is visually identified.

Step 2: These identified headings are then used to create a list containing the homepage URL and the URLs of valuable pages.

Step 3: For the constructed webgraph of the site, a vector of the PRs of all the nodes (pages) of the webgraph is computed. In this vector, only valuable nodes are assigned numbers starting from number 1. The homepage (being one of the valuable nodes) is assigned number 1.

Step 4: The vector is ordered in descending order of PRs.

Step 5: The list of valuable pages (Step 2) and the ordered PR vector (Step 4) are compared to identify any valuable page with low PR or non-valuable page with high PR, which are correctness criterion violations.

Step 6: A list of nodes violating the correctness criterion is created.

3.4. Control actions. The term "control action" means the addition and/or removal of new pages, documents and internal hyperlinks from a site. Obviously, control actions can differ very significantly — there are many of such actions and they are difficult to formalize. The authors believe that this is a separate area for further research.

Addition of new pages, documents and links to a site always depends on the behavior of the external environment of the site. A site usually reflects the activities of its owner. For example, if a newly employed lecturer has his profile page added in the website of the university, then links pointing to the pages of that lecturer's department and faculty, and

maybe also to the sports club (if the lecturer does sports), need to be added in this new profile page. This is what is meant by "behavior of the external environment".

It is difficult to formalize and model the behavior of the external environment. However, it is somewhat much easier to formalize and model the action of removing pages, documents and links from a website. By analyzing a particular site, one can identify the conceptual meaning of such removal. For example, if a lecturer is fired from the university, then links pointing from/to his profile page to/from other pages of the university website are simply deleted. If the lecturer had several pages (directory), then his directory and all links pointing from/to it are deleted. Besides, removal of a directory does not always mean that it has been totally eliminated. For example, the lecturer may have decided to create his own personal website instead of a section (directory) on the university's website. In this particular case, "removal" means "creating a new site and removing directory from the old site". This action is much easier to formalize.

The removal procedure is most obvious when it comes to empty pages, irrelevant and outdated documents, etc.

As shown in the above example, removal of some part of a site does not often imply physical removal. In many cases, it means transforming such part into an independent site with a unique domain name, and at the same time, retaining the hyperlinks (now external hyperlinks) linking that initial part to the parent site.

Such a process is natural for university web resources. Evolution of the web space of Saint Petersburg State University over 20 years of its existence was shown in [19]. The web space was initially composed of only the official university website and gradually grew to 280 independent sites of faculties, institutes and administrative divisions. Many of these bodies where previously sections in the official website of the university.

In this paper, therefore, control actions involve removing pages, documents and links from a website.

Understandably, the authors cannot delete a part of a real site since they are not professional web developers. However, they can experiment on the webgraph of the site in question to see the effect of such actions. In the simplest case, this involves removing some nodes and all their incident arcs.

A more challenging task is to remove some part of the graph. This means removing a certain number of nodes and their incident arcs.

In this paper, a part of a graph to be deleted is a sub-graph of the graph for which the root node is given. The remaining nodes are formed as follows:

Step 1: All the child nodes of the root node are selected.

Step 2: For all the child nodes selected in Step 1, all the nodes lying one level lower (the direct children of the child nodes selected in Step 1) are selected.

Step 3: This process is continued until leaf nodes are reached.

In web development, a part of a website to be deleted is considered a subdirectory, whose root is the page representing the selected node.

In subsequent experiments, only one type of control action is considered: removal of parts of graph. In this case, of particular interest is the question of whether the PageRank of the remaining valuable pages will improve or not. I should be noted that:

- there are many ways to change the structure of a site;
- this research is experimenting on only one of them — removal of "bad" valuable pages;

• the research is NOT interested in what happens to the "deleted" content; may be NEW sites are created from them or may be not;

• the research is interested in what would happen to the PageRank of the remaining (non-deleted) valuable pages. If the PR increases, then that's great, if it reduces but remains high, that's also good.

4. Experiments and results. The websites of faculties and colleges of three selected major universities in Nigeria, Russian Federation and USA were selected as the experimental samples. Experiments have clearly shown that the described approach is feasible. The websites are:

• http://engineering.unn.edu.ng — Faculty of Engineering, University of Nigeria, Nsukka;

• http://www.apmath.spbu.ru — Faculty of Applied Mathematics and Control Processes, Saint Petersburg State University, Russian Federation;

• http://cmns.umd.edu — College of Computer, Mathematical, and Natural Sciences, University of Maryland, USA.

These websites are characterized by different sizes and content diversity. They are broken down into sufficiently independent fragments. With the above main features and a number of others, all the steps involved in the approach can be conveniently presented through simple examples. The following abbreviations are used:

• **FOE**: Faculty of Engineering, University of Nigeria, Nsukka;

• **AMCP**: Faculty of Applied Mathematics and Control Processes, Saint Petersburg State University, Russian Federation;

• **CMNS**: College of Computer, Mathematical, and Natural Sciences, University of Maryland, USA.

For AMCP and CMNS experiments, RCCrawler scanned up to five levels deep — from 0 to 4. The reasoning behind this action is that pages below level 4 contribute little or nothing to the main features of a webgraph [20]. For FOE experiment, the crawler scanned all the levels because of the small size of the site (small page count). The FOE has only four levels.

In terms of graph-theoretical model, documents contained in websites are leaf nodes that do not have outgoing arcs. They have no significant influence on graph connectivity. Therefore, the authors restrict themselves only to html pages.

Taking into account the above said, the websites were scanned with the RCCrawler. The scanning results are shown in Table 1.

Table 1. **Scanning results**

Websites	Page count	Link count
http://engineering.unn.edu.ng (Faculty of Engineering, University of Nigeria, Nsukka)	79	1.680
http://www.apmath.spbu.ru (Faculty of Applied Mathematics and Control Processes, Saint Petersburg State University, Russian Federation)	3.812	242.202
http://cmns.umd.edu (College of Computer, Mathematical, and Natural Sciences, University of Maryland, USA)	478	9.296

Images of the webgraphs (in different scales) constructed in Gephi, are shown in Figure 3. The source (root) node of the graph is located approximately at the center of each figure. The remaining nodes are depicted depending on their degree of distance from the source node.

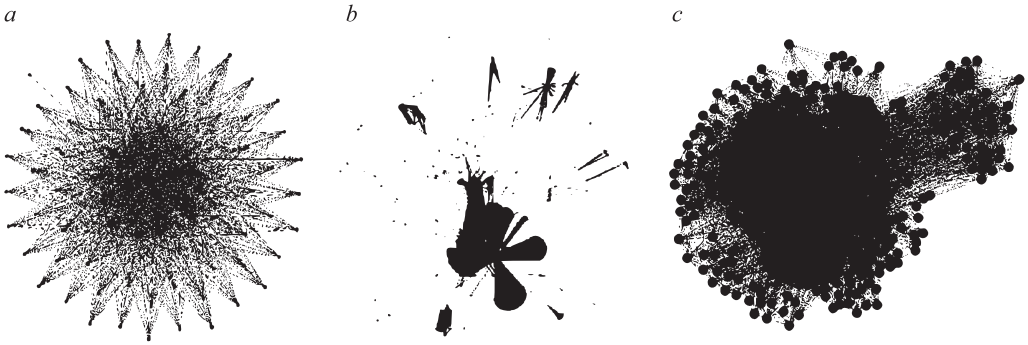

Figure 3. Webgraphs of FOE (*a*), AMCP (*b*) and CMNS (*c*)

A cursory look at the images shows that the FOE graph is symmetric in nature, the AMCP graph is incoherent, while the CMNS graph is compact.

4.1. Website of the Faculty of Engineering, University of Nigeria, Nsukka. The analysis starts with a visual and simple case for FOE. Table 2 is constructed in accordance with the procedure described in the section "Determining the Correctness Criterion for a Website Structure".

Table 2. **Details of the top pages of the FOE site, corresponding to the webgraph nodes**

Number assigned to valuable page	Section title	Webpage URL	PR	PR-reduce
1	Main	/	0.0586	0.0408
10	Alumni	/alumni/	0.0585	delete
no num	Administrative Building	/2015/11/12/slider1/	0.0358	delete
2	History	/about/history/	0.0358	0.0424
3	Key Officers	/about/key-officers/	0.0358	0.0424
4	Dean's Statement	/about/deans-statement/	0.0358	0.0424
5	Philosophy	/faculty/philosophy/	0.0358	0.0424
6	Admission Requirement	/faculty/admission-requirement/	0.0358	0.0424
7	Departments	/faculty/departments/	0.0358	0.0424
8	Staff Profiles	/faculty/staff-profiles/	0.0358	0.0424
9	External Resources	/external-resources/	0.0358	0.0424
11	Contact Us	contact-us/	0.0358	0.0424
12	Agric. Bio resources Engineering	/faculty/departments/agric-bioresources-engineering/	0.0358	0.0424
13	Civil Engineering	/faculty/departments/civil-engineering/	0.0358	0.0424
14	Electrical Engineering	/faculty/departments/electrical-engineering/	0.0358	0.0424
15	Electronic Engineering	/faculty/departments/electronic-engineering/	0.0358	0.0424
16	Mechanical Engineering	/faculty/departments/mechanical-engineering/	0.0358	0.0424
17	Metallurgical Materials Engineering	/faculty/departments/metallurgical-materials-engineering/	0.0358	0.0424

The first column of Table 2 contains the valuable page count of the site. The numbering is done starting from the left to the right of the sections of the main menu of the homepage

of the site engineering.unn.edu.ng, taking into account popup submenu. Figure 4 shows the image of this page with the added numbers of some sections. The names of sections are contained in the second column.

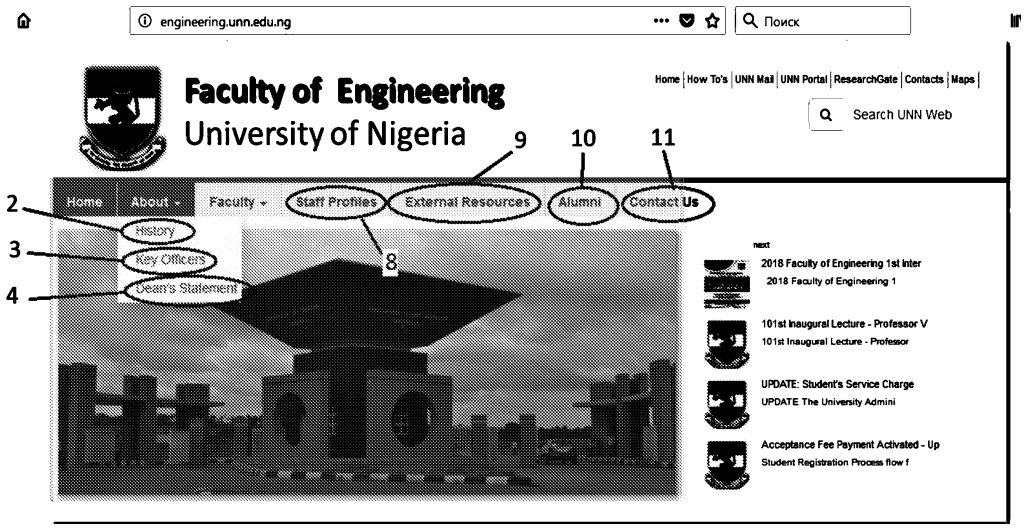

Figure 4. Part of the homepage of the FOE website engineering.unn.edu.ng

The URL column contains short addresses of the pages of the corresponding sections, without specifying the "http" protocol and the host http://engineering.unn.edu.ng. The table rows are ordered based on the descending order of PR values in the fourth column. These values are calculated (using Gephi) from a webgraph constructed using the obtained scan data.

In accordance with the procedure described in the section "Determining the Correctness Criterion for a Website Structure", the page corresponding to "Alumni" is found to have a very high PR compared to other valuable pages. Though not a valuable page (therefore no number is assigned to it), the "Administrative Building" page has a high PR.

The "Alumni" page was found to be empty and contains only one link to an equally empty directory. The "Administrative Building" page does not contain anything new, except the photo of the administrative building. Deleting the "Alumni", "Alumni/cgi-bin" and "Administrative Building" pages are the possible control actions in this case. The above action is simulated on a webgraph — the nodes corresponding to the "Alumni", "Alumni/cgi-bin", and "Administrative Building" pages and all their incident arcs are deleted from the webgraph. The "PR-reduce" values placed in the fifth column of Table 2 represent the PR values computed after control actions have been implemented.

Analysis shows that by removing two empty and one less informative page from the site, the PR value of almost all the valuable pages increases by 18 %.

Curiously, the PR value of the homepage decreases by 14 %. Theoretical investigations into the structure of the websites have shown that such a difference — where the PR value of level 1 pages is higher than that of level 0 — is not abnormal [21]. This claim is verified later in the CMNS experiment.

4.2. Website of the Faculty of Applied Mathematics and Control Processes, Saint Petersburg State University, Russian Federation. Figure 5 shows the homepage of the website (Russian version) of the Faculty of AMCP. The sections marked with numbers and highlighted in ovals are the valuable pages of the website. The "Section title" column is added in the comparison table (Table 3). This column contains Russian-to-English translated page names. As before, the "http" protocol and host "http://www.apmath.spbu.ru" are omitted in the "Webpage URL" column.

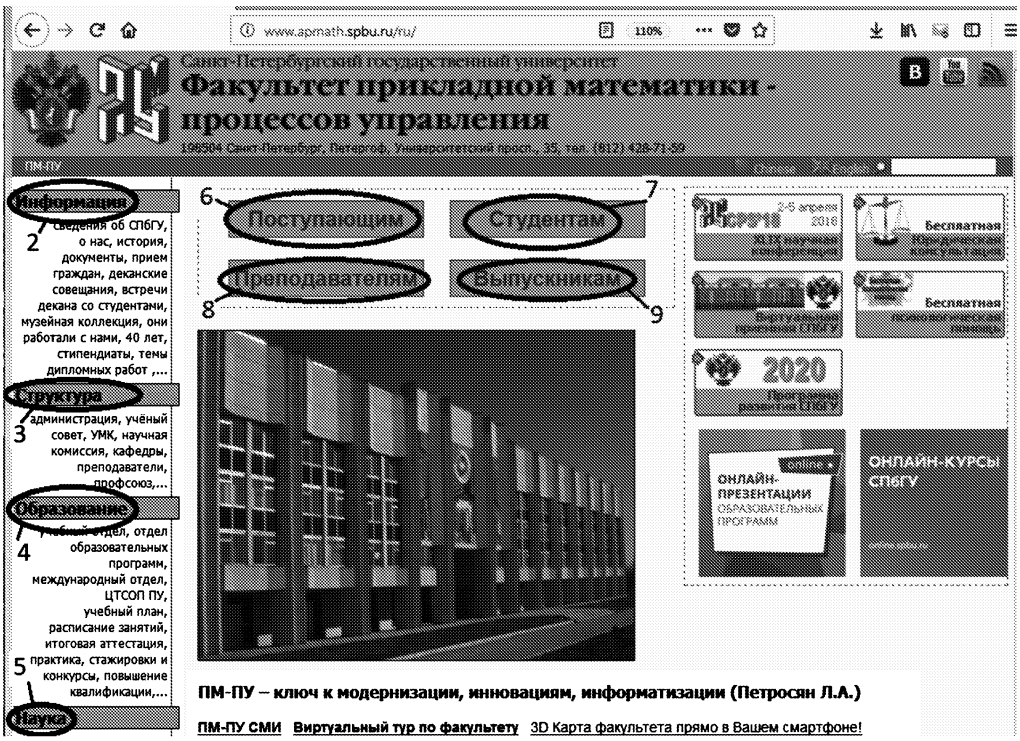

Figure 5. Part of the homepage of the AMCP website apmath.spbu.ru

In Table 3, part of the pages had to be skipped because valuable page "Science" occupies only 25th position in the ordered PR vector, while the "Information" page occupies 54th position. This means most of the valuable pages have very low PRs. It is not really a good sign when the PR of the "Web feed" page is six times higher than the PR of the Home page. The following actions are performed on the webgraph:

- the "Other" node is removed;
- all the nodes and arcs that make up the directory (for the site), whose root corresponds to the "Other" node are removed;
- all the arcs incident to the removed nodes are removed.

After these operations, the sub-graph deleted contains 1.914 nodes and 97.892 arcs. This is essentially half of the entire webgraph of the site.

Naturally, it can be assumed that "deletion" or "removal" actually means that the directory being deleted must be transformed into an independent site, for example,

Table 3. **Details of some pages of the AMCP site, representing the nodes of the webgraph**

Number assigned to valuable page	Section title (Russian)	Section title (translated to English)	Webpage URL	PR	PR-reduce
no num	с	Web feed	/ru/rss.php/	0.09508	0.06677
1	Главная	Main	/ru/	0.01537	0.013934
no num	Разное	Other	/ru/misc/	0.01284	delete
no num	Фотоальбом	Photo gallery	/ru/misc/album/2015/	0.00792	delete
no num	Фотоальбом	Photo gallery	/ru/misc/album/2010/	0.00792	delete
...	...	...	...	...	...
5	Наука	Science	/ru/research/	0.00296	0.005513
3	Структура	Structure	/ru/structure/	0.00249	0.005513
...	...	...	...	...	...
2	Информация	Information	/ru/info/	0.00245	0.006772
4	Образование	Education	/ru/education/	0.00245	0.005513
6	Поступающим	Admission	/ru/admission/r.html	0.00245	0.005513
9	Выпускникам	For alumni	/ru/alumni/	0.00245	0.005513
8	Преподавателям	For teachers	/ru/forstaff.html	0.00018	0.005513
7	Студентам	For students	/ru/forstudents.html	0.00018	0.005513

http://other.apmath.spbu.ru, while retaining the hyperlinks linking it to the parent site apmath. spbu.ru

The reduced graph contains 1.898 nodes and 14.4373 arcs. Its PR-reduce values are computed and placed in the sixth column of Table 3.

As a result, the position of valuable pages is significantly improved: the homepage now ranks first, the "Information" page is second, then all the other valuable pages follow, skipping only the "Web feed" page.

4.3. Website of the College of Computer, Mathematical, and Natural Sciences, University of Maryland, USA. The upper part of the homepage of the CMNS website is shown in Figure 6. The valuable sections are numbered and highlighted with ovals.

As can be seen from Table 4, the valuable pages of the site occupy nine first positions. From this, it can be concluded that the criterion for the correctness of the site structure is not violated and no control actions are required.

Table 4. **Details of the pages of the CMNS site with the highest PR values**

Number assigned to valuable page	Section title	Webpage URL	PR
1	Main	/	0.0448
2	About CMNS	/about-cmns	0.0458
3	Departments	/departments	0.0458
4	Research	/research	0.0458
5	News&Events	/news-events/news	0.0458
6	Undergraduate	/undergraduate	0.0458
7	Graduate	/graduate	0.0458
8	Alumni&friends	/alumni-friends	0.0458
9	Faculty&staff	/faculty-staff	0.0458
no num	Odyssey Magazine	/news-events/news/odyssey-magazine	0.0458
no num	Contact Us	/contact-us	0.0458
no num	Careers&Recruiting	/careers-recruiting	0.0458

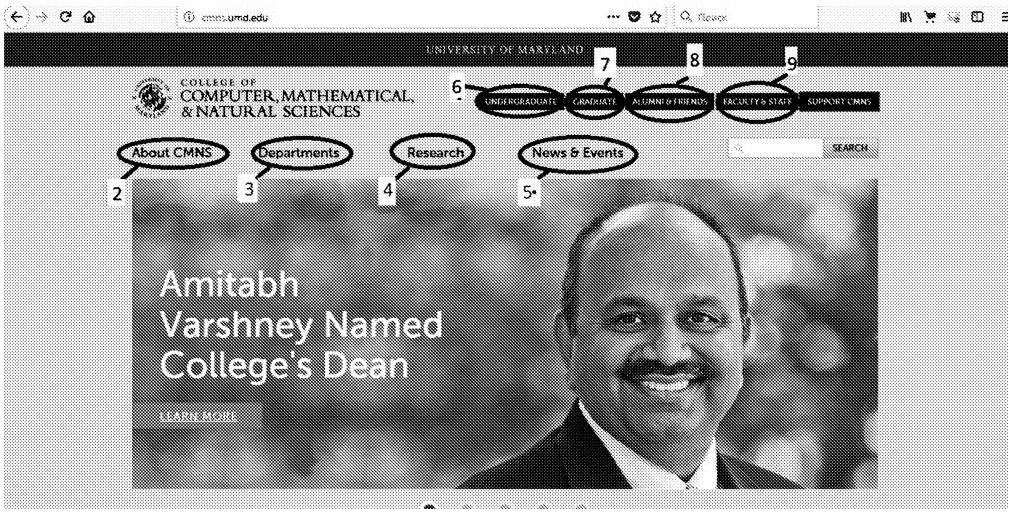

Figure 6. Part of the homepage of the CMNS website cmns.umd.edu

5. Discussion and recommendations. Based on experiments carried out, some conclusions were arrived at and recommendations were made.

The FOE website is still at the initial stage of development. The website has a number of pages that are not currently maintained. Such pages ought to be removed from the site. A preliminary plan for boosting the website structure should be developed. The plan should include converting some sections of this site into standalone (independent) sites, especially if such sections have sub-sections.

The structure of the AMCP website requires radical changes. First, the "Other" section is unrelated to academic activities, but is apparently important for the activities of the faculty. Therefore, it can be reorganized into an independent site. Then, as in the case above, a plan for reorganizing certain sections into independent sites should be developed.

Reorganizing certain sections into independent sites is aimed at ensuring that valuable pages (that can be accessed directly from the homepage) have the highest significance (PageRank) among all the pages of the site. This is achieved by separating into independent sites those sections whose root pages have high PageRank, but are not considered as being valuable pages by the site developer. If the developer had considered them as valuable pages, he would have placed links pointing to them from the homepage (i. e. such pages would have had direct access from the homepage of the site). The CMNS website seems to be a well-balanced and structured site and does not require major control actions.

6. Conclusion. A formalized procedure for website analysis was described. With a graph-theoretic model, this procedure allowed to examine the structure of a website, evaluate the PageRank values of its pages, and model the effects of control actions — removal of individual localized components from the site.

Step-by-step execution of the procedure was demonstrated in real-life examples: the website of the Faculty of Engineering, University of Nigeria, Nsukka, the website of the Faculty of Applied Mathematics and Control Processes, Saint Petersburg State University, Russian Federation, and the website of the College of Computer, Mathematical, and Natural Sciences, University of Maryland, USA. Based on the modeled effects of control actions, the authors provide guidelines on how to alter site structure in order to enhance it.

Experiments revealed that the structures of the FOE and AMCP websites need some changes. It was established that the CMNS website does not require such changes because its structure meets the proposed criterion.

It was found that lack of clearly defined regulations and development plans is one of the major obstacles towards formalizing an efficient decision-making model for a website structure.

This paper has clearly demonstrated that the proposed method has a high level of generality. It can be a vital tool in exploring many other academic websites.

Another benefit of the proposed approach is that any free graph/network software license, such as exploration software Gephi, can be used by a researcher to collect, analyze, process and visualize webometrics data with regards to similar research. Consequently, the proposed approach has a very wide range of potential users.

References

1. Björneborn L., Ingwersen P. Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, 2004, vol. 55(14), pp. 1216–1227.

2. Thelwall M. A history of webometrics. *Bulletin of the American Society for Information Science and Technology*, 2012, vol. 38(6), pp. 18–23.

3. Broder A., Kumar R., Maghoull F., Raghavan P., Rajagopalan S., Stata R., Tomkins A., Wiener J. Graph structure in the web. *Journal of Computer Networks*, 2000, vol. 33(1–6), pp. 309–320.

4. Ortega J. L., Aguillo I. F. Visualization of the Nordic Academic Web: Link analysis using social network tools. *Information Processing & Management*, 2008, vol. 44 (4), pp. 1624–1633.

5. Thelwall M., Harries G. The connection between the research of a university & counts of links to its web pages: An investigation based upon a classification of the relationships of pages to the research of the host university. *Journal of the American Society for Information Science and Technology*, 2003, vol. 54(7), pp. 593–699.

6. Babak A., Babak S. Graph theory application and web page ranking for website link structure improvement. *Behavior & Information Technology*, 2009, vol. 28 (1), pp. 63–72. https://doi.org/10.1080/01449290701840948

7. Liu Y., Ma Z. M., Zhou C. Web Markov skeleton processes and their applications. *Tohoku Mathematical Journal*, 2011, vol. 63, pp. 665–695.

8. Shokin Y. I., Klimenko O. A., Rychkova E. V., Shabalnikov I. V. Website rating for scientific and research organizations of the Siberian Branch of Russian Academy of Sciences. *Computational Technologies.* 2008, vol. 13 (3), pp. 128–135.

9. Dehmer M., Dobrynin A. A., Konstantinova E. V., Vesnin A. Y., Klimenko O. A, Shokin Y. I., Rychkova E. V., Medvedev A. N. Analysis of webspaces of the Siberian Branch of the Russian Academy of Sciences and the Fraunhofer—Gesellschaft. *Information Technology in Industry.* 2018, vol. 6 (1), pp. 1–6.

10. Meusel R., Vigna S., Lehmberg O., Bizer C. The graph structure in the web — analyzed on different aggregation levels. *The Journal of Web Science*, 2015, vol. 1, pp. 33–47. https://doi.org/10.1561/106.00000003

11. Pant G., Srinivasan P., Menczer F. Crawling the web. *Web dynamics.* Eds by M. Levene, A. Poulovassilis. Berlin, Springer Publ., 2004, pp. 153–178.

12. Yadav M., Goyal N. Comparison of open source crawlers — a review. *International Journal of Scientific & Engineering Research*, 2015, vol. 6 (9), pp. 1544–1551.

13. Pechnikov A. A., Lankin A. V. Development of a program for collecting data on the structure of websites. *Proceedings of the Karelian Research Center of the Russian Academy of Sciences*, 2016, vol. 8, pp. 81–90. https://doi.org/10.17076/mat381

14. Bar-Yossef Z., Keidar I., Schonfeld U. Do not crawl in the DUST: Different URLs with similar text. *Proceedings of the 16th International World-Wide Web Conference*, 2007, pp. 111–120. https://doi.org/10.1145/1242572.1242588

15. Cormen T. H., Leiserson C. E., Rivest R. L., Stein C. Breadth-first search. *Introduction to Algorithms.* 2nd ed. Cambridge, Massachusetts, MIT Press and McGraw-Hill Publ., 2001, pp. 531–539.

16. *The Open Graph Viz Platform.* Available at: https://gephi.org (accessed: 10.01.2019).

17. Brin S., Page L. The anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 1998, vol. 30 (1–7), pp. 107–117.

18. Pandurangan G., Raghavan P., Upfal E. Using PageRank to characterize web structure.

Proceedings of the 8th Annual International Conference: Computing and Combinatorics, 2000, vol. 2387, pp. 330–339.

19. Pechnikov A. A. Web resources of the Russian university: Self-organization or administrative impact? *Electronic Libraries*, 2015, vol. 18 (6), pp. 277–295.

20. Baeza-Yates R., Castillo C. Crawling the infinite web: Five levels are enough. *Proceedings of the Third Workshop on Web Graphs (WAW) of the Lecture Notes in Computer Science. Algorithms and Models for the WebGraph*, 2004, vol. 3243, pp. 156–167.

21. Pechnikov A. A. On evaluating the value of the pages of a website. *Informatization of Education and Science*, 2015, vol. 4 (28), pp. 28–40.

Received: March 03, 2019.
Accepted: June 06, 2019.

Author's information:

Anthony M. Nwohiri — PhD; anwohiri@unilag.edu.ng

Andrey A. Pechnikov — Dr. Sci. in Technics; pechnikov@krc.karelia.ru

Применение вебометрических методов анализа структуры сайта для улучшения его структуры по критерию ценности страниц

Э. М. Нвохири[1], *А. А. Печников*[2]

[1] Университет Лагоса, Нигерия, 101017, Лагос, Университи Роуд, Акока-Яба
[2] Институт прикладных математических исследований Карельского научного центра Российской академии наук, Российская Федерация, 185910, Петрозаводск, ул. Пушкинская, 11

Для цитирования: *Nwohiri A. M., Pechnikov A. A.* Application of webometrics methods for analysis and enhancement of academic site structure based on page value criterion // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2019. Т. 15. Вып. 3. С. 337–352. https://doi.org/10.21638/11702/spbu10.2019.304 (In English)

Описана формализованная процедура исследования веб-сайта вебометрическими методами, включающая сбор данных о его структуре, построение и исследование веб-графа сайта, определение критерия «правильности» структуры сайта, определение управляющих воздействий для улучшения структуры сайта в смысле заданного критерия, проверка критерия на реальных примерах и разработка рекомендаций по улучшению структуры сайта. В качестве критерия оценки значимости страниц используется их *Web PageRank* (*PR*). Ценность страницы определяется по наличию (или отсутствию) ссылки на нее на главной странице сайта. Страница считается «ценной», если такая ссылка имеется. Критерий «правильности» структуры сайта определяется так: ценные страницы сайта должны иметь наивысший PR среди всех страниц сайта (главная страница считается ценной по умолчанию). Как управляющее воздействие рассматривается удаление директорий (с выделением их в самостоятельные сайты), имеющих в качестве корня страницы с высоким PR, но не являющиеся ценными. Проведенные эксперименты демонстрируются на примере факультетских сайтов крупных университетов Нигерии, России и США и во всех случаях показывают свои применимость и состоятельность.

Ключевые слова: веб-сайт, график, PageRank, университеты, интеллектуальный анализ данных, структура веб-сайта, извлечение веб-данных, веб-майнинг, URL.

Контактная информация:

Нвохири Энтони М. — PhD; anwohiri@unilag.edu.ng

Печников Андрей Анатольевич — д-р техн. наук; pechnikov@krc.karelia.ru